The conditioning of linear boundary value problems

Mattheij, R.M.M.

Published in:
SIAM Journal on Numerical Analysis

DOI:
10.1137/0719070

Published: 01/01/1982

Document Version
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
THE CONDITIONING OF LINEAR BOUNDARY VALUE PROBLEMS*

R. M. M. MATHIEIJ†

Abstract. Investigation is made into the sensitivity of solutions of linear boundary value problems to perturbations of the boundary condition. We derive a useful quantity to decide for well or ill conditioning. From this it is deduced which kind of requirements the boundary conditions should meet in order to have a well conditioned problem. It is shown that this quantity can fruitfully be used to explain why, e.g., the multiple shooting technique is stable (in contrast to the single shooting one) for certain well posed problems having solutions with different growth behavior. The results are sustained by a number of examples.

1. Introduction. The study of boundary value problems (BVP) of ODE has produced a large number of numerical methods. Curiously enough, the conditioning of the problem as such has received much less attention. In fact, one would like to have something like condition numbers used in linear algebra. Obviously inspired by this, e.g., in [5] an analysis is given to decide for the conditioning of shooting methods (see also [4]). However, since shooting essentially is an initial value method, this result is not really what we are aiming at. Indeed, we would rather have some criterion that indicates the nature of the problem, not of some method. (As is well known, shooting may not be stable at all, despite physical arguments that the problem is well conditioned, cf. e.g. [7], [8].)

It is the intention of this paper to discuss the inherent (in-)stability of a BVP. This is done by introducing condition numbers that indicate the possible amplification of perturbations in an absolute sense (this is induced by the fact that one is usually interested in absolute accuracy). From heuristic arguments, it is natural to ask whether the number of increasing and/or decreasing solutions (if present) are related to the number of boundary conditions imposed at each of the boundary points. This question is also of importance, since several algorithms for problems with linear boundary conditions (cf. [15]) essentially make use of knowledge about the number of independent relations at the initial point, say. Therefore a substantial part of this paper is devoted to finding out the relationship between the conditioning and the rank of the boundary conditions. The usefulness of this relationship is demonstrated by Example 6.5.

Although the condition numbers in principle indicate by how much any possible errors in the boundary conditions may be amplified, it turns out that they also play an important role in estimating the global error due to other perturbations. This is analyzed in [11]. An intriguing question is whether the stability of certain algorithms, like multiple shooting, is related to the conditioning of the problem. Therefore we give a stability analysis of this algorithm and show that the condition number is an important quantity indeed in estimating the global error. We believe that this aspect of our results is new and may give a valuable contribution to the understanding of multiple shooting.

The paper is built up as follows: First we give in § 2 a more detailed problem setting and some definitions. Then in § 3 we introduce the notion of the condition number of a problem. The relation of the conditioning to the boundary conditions is considered in § 4. In § 5 we investigate the stability of the multiple shooting algorithm. Finally, we illustrate the analysis of §§ 4, 5 by a number of examples in § 6.

* Received by the editors March 26, 1980, and in revised form November 4, 1981.
† Mathematisch Instituut, Katholieke Universiteit, Toernooiveld, 6525 ED Nijmegen, the Netherlands.

963
2. The problem setting and some conventions

2.1. Norms. The norm \( \| \cdot \| \) that will be used is assumed to be a Hölder norm. Only in some special cases do we use the explicit notation \( \| \cdot \|_p \) (for some \( p \)). The induced matrix norm will also be denoted by \( \| \cdot \| \). Besides, we shall use the \( \text{glb} \) of a matrix \( A \), defined as

\[
\text{glb}(A) = \min_{x \neq 0} \frac{\|Ax\|}{\|x\|}.
\]

If \( A^{-1} \) exists, then \( \text{glb}(A) = \|A^{-1}\|^{-1} \).

Remark 2.2. If \( A \) is column equilibrated, say all columns having unit length, then \( \text{glb}(A) \) will only be small if any of the columns of \( A \) makes a small angle with the subspace spanned by the rest. More precisely, if \( A = (a_1 \cdots |a_n) \) and \( \theta \) is the smallest angle between any column of \( A \) and the subspace spanned by the rest, then \( \text{glb}_2(A) \geq \sin \theta \min_{i=1, \ldots, n} \|a_i\|_2 \) (cf. [9, § 4]).

2.2. The ODE and its solutions. Consider the ODE

\[
\frac{d}{dt} x(t) = A(t)x(t) + f(t), \quad t \in [\alpha, \beta] \subseteq \mathbb{R}
\]

where \( A \) is an \( n \times n \)-matrix function and \( f \) an \( n \)-vector function. Let \( \Phi \) denote a fundamental solution, i.e., an \( n \times n \)-matrix solution of the homogeneous part

\[
\frac{d}{dt} \Phi(t) = A(t)\Phi(t).
\]

An interesting class of BVP involves ODE with both increasing solutions and non-increasing solutions. For such ODE it is not restrictive to assume that this dichotomy is reflected in the partitioning of the matrices \( \Phi(t) \). Inspired by this we thus have:

Assumption 2.5. Let \( \Phi(t) \) be written as \( (\phi^1(t)|\cdots|\phi^n(t)) \). Let there exist integers \( k, l, 0 \leq k, l \leq n \), and moreover real numbers \( \lambda > 0 \), \( \mu < 0 \) and positive constants \( \gamma_1 \), \( \gamma_2 \) and \( \gamma_3 \), such that for all \( t, \tau \in [\alpha, \beta] \) with \( \tau \geq t \) there hold

(i) \[
\frac{\|\phi^j(t)\|}{\|\phi^j(\tau)\|} \leq \gamma_1 e^{-\lambda(\tau-t)}, \quad j \leq k,
\]

(ii) \[
\frac{\|\phi^j(\tau)\|}{\|\phi^j(t)\|} \leq \gamma_2 e^{\mu(\tau-t)}, \quad j \geq n-l+1,
\]

and, if \( k < n-l \), also

(iii) \[
\frac{\|\phi^j(\tau)\|}{\|\phi^j(t)\|} \leq \gamma_3, \quad k+1 \leq j \leq n-l.
\]

Such a type of fundamental solution \( \Phi \) naturally arises in the constant coefficient case, where \( k \) corresponds to the number of eigenvalues with a positive real part, and \( l \) to the number of eigenvalues with a negative real part.

It is convenient for later use to partition the matrices \( \Phi(t) \) into blocks of columns as is induced by Assumption 2.5, viz

\[
\Phi(t) = [\Phi^1(t)\Phi^2(t)\Phi^3(t)],
\]

where \( [\Phi^2(t)] \) may be nonexistent if there are only increasing and decreasing solutions, i.e., \( k+l = n \), etc.

There are of course also other classes of solution spaces (e.g., with solutions that decrease on a certain subinterval, but increase outside that interval). Nevertheless we
think that the class of ODE with solutions as indicated in Assumption 2.5 enables us to give a fairly complete description of problems concerning the conditioning. At the end of § 4 we return to this question.

2.3. The boundary conditions. The general linear boundary condition (BC) for the solution $x$ can be written as

$$(2.7) \quad M_\alpha x(\alpha) + M_\beta x(\beta) = b,$$

where $M_\alpha, M_\beta$ are $n \times n$ matrices and $b$ is an $n$-vector. For a solution of (2.3) satisfying (2.7) we have the well known results (cf. [8]):

Property 2.8. There exists a unique solution of (2.3) and (2.7) if and only if $M_\alpha \Phi(\alpha) + M_\beta \Phi(\beta)$ is nonsingular.

Property 2.9. A solution of (2.3) and (2.7) is unique only if $\text{rank } [M_\alpha \quad M_\beta] = n$.

Quite often one has separated BC, being of the form

a) $$S_\alpha x(\alpha) = b_\alpha,$$

b) $$S_\beta x(\beta) = b_\beta,$$

where $S_\alpha, S_\beta$ are (not necessarily square) matrices and $b_\alpha, b_\beta$ vectors of suitable dimensions. We will always assume that $[S_\alpha \quad S_\beta]$ is an $n \times n$ matrix. By supplementing $S_\alpha$ and $S_\beta$ with zeros we can regard (2.10) as a special case of (2.7). We have

Property 2.11. Let $[S_\alpha \quad S_\beta]$ be an $n \times n$-matrix. Then there exists a unique solution of (2.3) and (2.10) if and only if

$$\begin{bmatrix} S_\alpha \Phi(\alpha) \\ S_\beta \Phi(\beta) \end{bmatrix}$$

is nonsingular.

For the sake of convenience we will assume that the matrices $M_\alpha$ and $M_\beta$ (whether or not arising from separated BC) are scaled such that

$$(2.12) \quad \max (\|M_\alpha\|, \|M_\beta\|) = 1.$$

If the solution $x$ is unique, then a BVP is called well posed. From Property 2.8 we see that for a well posed problem $[M_\alpha \Phi(\alpha) + M_\beta \Phi(\beta)]^{-1}$ must exist. In the next section we will see that this existence does not imply well conditioning at all.

3. Condition numbers. In computing a solution $x$ by some method, rounding errors and possibly truncation errors are made. Their effects on the computed solution depend on the problem and on the method which is used. It turns out that the question of how such local errors affect the global error is strongly connected with the fundamental problem of how perturbations of the BC influence the solution. It is the purpose of this section to analyze this sensitivity of the solution with respect to the BC. This will eventually lead to the notion of condition number. We first give an example.

Example 3.1. Consider the ODE

$$\frac{d^2 u}{dt^2} = 100u,$$

or in vector notation, with $x = (u, du/dt)^T$,

$$(3.3) \quad \frac{dx}{dt} = \begin{bmatrix} 0 & 1 \\ 100 & 0 \end{bmatrix} x.$$
Obviously a fundamental solution $\Phi$ is given by
\begin{equation}
\Phi(t) = \begin{bmatrix}
e^{10t} & e^{-10t} \\
10e^{10t} & -10e^{-10t}
\end{bmatrix}.
\end{equation}

Consider the BC
\begin{equation}
\begin{aligned}
\frac{du^\varepsilon}{dt}(0) + \gamma \frac{du^\varepsilon}{dt}(0) &= 1 + \varepsilon_1, \\
\frac{du^\varepsilon}{dt}(1) + \delta \frac{du^\varepsilon}{dt}(1) &= 1 + \varepsilon_2.
\end{aligned}
\end{equation}

Define in an obvious way $x^\varepsilon = (u^\varepsilon, du^\varepsilon/dt)^T$ (where the superscript $\varepsilon$ denotes the dependence on the vector $(\varepsilon_1, \varepsilon_2)^T$). Then we have in matrix notation
\begin{equation}
\begin{bmatrix}
1 & \gamma \\
0 & 0
\end{bmatrix} x^\varepsilon(0) + \begin{bmatrix}
0 & 0 \\
1 & \delta
\end{bmatrix} x^\varepsilon(1) = \begin{bmatrix}
1 + \varepsilon_1 \\
1 + \varepsilon_2
\end{bmatrix}.
\end{equation}

Suppose we are interested in the solution $x^0$, but for some reason we have a perturbed BC (3.6) with $\varepsilon \neq 0$. Then the error $z := x^\varepsilon - x^0$ satisfies the BC
\begin{equation}
\begin{bmatrix}
1 & \gamma \\
0 & 0
\end{bmatrix} z(0) + \begin{bmatrix}
0 & 0 \\
1 & \delta
\end{bmatrix} z(1) = \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2
\end{bmatrix}.
\end{equation}

Since $\Phi$ is a fundamental solution of (3.3) there exists a vector $v \in \mathbb{R}^2$ such that
\begin{equation}
z = \Phi v.
\end{equation}

Substituting this in (3.7) we find:
\begin{equation}
v = \kappa(\gamma, \delta) \begin{bmatrix}
e^{-10(1-10\delta)} & 10\gamma - 1 \\
e^{+10(1+10\delta)} & 10\gamma + 1
\end{bmatrix} \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2
\end{bmatrix},
\end{equation}

where $[\kappa(\gamma, \delta)]^{-1} = [e^{-10(1-10\delta)(1+10\gamma)} - e^{10(1-10\gamma)(1+10\delta)}]$. Hence for the error $u^\varepsilon(t) - u^0(t)$ we obtain in particular
\begin{equation}
u(t) - u^0(t) = \kappa(\gamma, \delta)[(e^{10t(1-10\delta)} - e^{-10(t-1)(10\delta + 1)})\varepsilon_1 \\
+ e^{10t(10\gamma - 1)} + e^{-10t(10\gamma + 1)}\varepsilon_2].
\end{equation}

If $\gamma = \delta = 0$, then we see that $\kappa(0, 0) = e^{-10}$, i.e., the error $|u^\varepsilon(t) - u^0(t)|$ nowhere exceeds $|\varepsilon_1| + |\varepsilon_2|$ by very much. If, however, $\gamma = 1, 0$ or $\delta = -1, 0$ then $|u^\varepsilon(t) - u^0(t)|$ may be a factor $e^{20}$ larger (for, e.g., $\delta$ and $\gamma$, resp. $= 0$). The former case apparently is "well conditioned," whereas the latter turns out to be "ill conditioned."

Remark 3.11. It is well known that an initial value problem or a terminal value problem for (3.2) is not well conditioned. Note, however, that the errors are amplified then by a factor $e^{-10}$ at most.

Remark 3.12. Since (3.7) and (3.8) give a linear equation for $v$, $Qv = [e^s]$ say, one might have the feeling that the condition number $\|Q\|\|Q^{-1}\|$ of this system is indicative of well conditioning (cf. [18, p. 191]). However, if $\gamma = \delta = 0$, this number is $\approx e^{10}$ and if, e.g., $\gamma = 1, 0$, $\delta = 0$ this number is $\approx e^{30}$. Although this last number clearly indicates that the latter case is more ill conditioned than the former, we may conclude that it is not the decisive criterion.

The reason that the quantity $\|Q\|\|Q^{-1}\|$ in Remark 3.12 is not the decisive criterion is, of course, that we are interested in $\Phi(t)v$ and not in $v$. This induces the following:

Definition 3.13. The condition number of the BVP is defined as
\begin{equation}
\mathcal{CN} := \max_{t \in [\alpha, \beta]} \|\Phi(t)[M_\alpha \Phi(\alpha) + M_\beta \Phi(\beta)]^{-1}\|.
\end{equation}
Remark 3.14. The number $\mathcal{CN}$ is in fact independent of the choice of the fundamental solution. This follows from the observation that for any other fundamental solution $\tilde{\Phi}$, say, there exists a nonsingular matrix $P$, say, such that $\forall \tilde{\Phi}(t) = \Phi(t)P$.

Suppose $\|\| = \|\|$ in Example 3.1. Then from (3.4) and (3.9) it can easily be seen that for $\gamma = \delta = 0$, the value of $\mathcal{CN} \approx 11$ and for $\gamma = \frac{1}{10}$ and $\delta = 0$ this value will be $=11 \cdot e^{20}$. Hence $\mathcal{CN}$ gives a proper description of the actual error amplification.

Below we show more generally that $\mathcal{CN}$ is the right criterion to indicate possible error amplification of perturbations of the BC, indeed. To this end, let $\delta M_a, \delta M_B$ and $\delta b$ be perturbations of $M_a, M_B$ and $b$, respectively. Let $\tilde{x}$ be a solution of (2.3) and the BC

\begin{equation}
(M_a + \delta M_a)\tilde{x}(\alpha) + (M_B + \delta M_B)\tilde{x}(\beta) = b + \delta b.
\end{equation}

Then there certainly exists a fixed vector $v$, such that

\begin{equation}
x(t) - \tilde{x}(t) = \Phi(t)v \quad \text{for all } t \in [\alpha, \beta].
\end{equation}

For ease of writing we introduce the matrices $Q$ and $\delta Q$ defined by

\begin{align*}
\text{a) } Q &:= M_a \Phi(\alpha) + M_B \Phi(\beta), \\
\text{b) } \delta Q &:= \delta M_a \Phi(\alpha) + \delta M_B \Phi(\beta).
\end{align*}

First we prove that $\|\delta QQ^{-1}\|$ can be estimated in terms of $\mathcal{CN}$ and the perturbations:

**Lemma 3.18.** $\|\delta QQ^{-1}\| \leq (\|\delta M_a\| + \|\delta M_B\|)\mathcal{CN}$.  

**Proof.** $\|\delta QQ^{-1}\| \leq \|\delta M_a\|\|\Phi(\alpha)Q^{-1}\| + \|\delta M_B\|\|\Phi(\beta)Q^{-1}\|$. □

We now have:

**Theorem 3.19.** Let $\varepsilon$ be such that $0 < \varepsilon < 1/2\mathcal{CN}$. If the perturbations are such that $\max(\|\delta M_a\|, \|\delta M_B\|, \|\delta b\|) = \delta \leq \varepsilon$, then the solution $\tilde{x}$ of (3.15) and (2.3) satisfies

\[
\max_{t \in [\alpha, \beta]} \|x(t) - \tilde{x}(t)\| \leq \frac{\mathcal{CN}(1 + \|x(\alpha)\| + \|x(\beta)\|)}{1 + 2\mathcal{CN} \delta}.
\]

Moreover, given any such perturbation bound $\delta \leq \varepsilon$, there always exists a solution $\tilde{x}$ of (3.15) and (2.3) such that

\[
\frac{\mathcal{CN}(1 + \|x(\alpha)\| + \|x(\beta)\|)}{1 + 2\mathcal{CN} \delta} \leq \max_{t \in [\alpha, \beta]} \|x(t) - \tilde{x}(t)\| \leq \frac{\mathcal{CN}(1 + \|x(\alpha)\| + \|x(\beta)\|)}{1 - 2\mathcal{CN} \delta}.
\]

**Proof.** From (2.7), (3.15), (3.16) and (3.17) we obtain

\[
(Q + \delta Q)v = w := \delta b - \delta M_a x(\alpha) - \delta M_B x(\beta).
\]

Hence

\begin{align*}
\text{(a) } v &= Q^{-1}(I + \delta QQ^{-1})^{-1}w. \\
\text{(b) } \|w\| &= \delta (1 + \|x(\alpha)\| + \|x(\beta)\|).
\end{align*}

Substitution of (a) into (3.16) and taking norms gives $\|\Phi(t)v\| \leq \mathcal{CN}(1/(1 - \|\delta QQ^{-1}\|))\|w\|$, from this the right-hand side inequalities easily follow.

To show the left-hand side inequality, we first note that for any $\delta b$ with $\|\delta b\| = \delta$, there exist $\delta M_a, \delta M_B$ with $\|\delta M_a\| = \|\delta M_B\| = \delta$, $\|\delta M_a x(\alpha)\| = \delta \|x(\alpha)\|$ and $\|\delta M_B x(\beta)\| = \delta \|x(\beta)\|$, and moreover $\delta M_a x(\alpha)$ and $\delta M_B x(\beta)$ are equal to $-\delta b$, apart from positive constants (this follows from [16, § 6]). For a vector $w$ as above, composed of such vectors, there apparently holds

\[
\|w\| = \delta (1 + \|x(\alpha)\| + \|x(\beta)\|).
\]
Now let the vector $u$ and the point $\tilde{x} \in [\alpha, \beta]$ be such that

$$(c) \| \Phi(\tilde{x})Q^{-1}u\| = \mathcal{C}_N\|u\|.$$  

Then we choose $w(=\delta b - \delta M_\alpha x(\alpha) - \delta M_\beta x(\beta))$ such that $[I + \delta QQ^{-1}]^{-1}w$ has the direction of the (maximizing) vector $u$ in (c). On account of this choice and of (b) we thus find that there exists a solution $\tilde{x}$ such that for $\Phi(\tilde{x})v = x(\tilde{x}) - x(\tilde{x}))$ the following bound holds:

$$\|\Phi(\tilde{x})v\| = \|\Phi(\tilde{x})Q^{-1}\|\|(I + \delta QQ^{-1})^{-1}w\| \leq \frac{\mathcal{C}_N\delta(\|x(\alpha)\| + \|x(\beta)\|)}{1 - 2\mathcal{C}_N\delta}. \quad \Box$$

In order to have a more workable quantity for $\mathcal{C}_N$ we may estimate it by

$$\mathcal{C}_N \leq \max_{\alpha, \beta} \|\Phi(\alpha)\|\|Q^{-1}\|.$$ 

Such an estimate appears, e.g., in [5]. However, if we choose $\Phi(\alpha) = I$, as is done there, then $\max \|\Phi(\alpha)\|$ threatens to be the dominant quantity in (3.20). For Example 3.1 this would mean that we would have an estimate that is a factor $\epsilon^{10}$ too large (it indicates the conditioning of some algorithms though, cf. (5.9)). We will not have those problems if we scale $\Phi$ appropriately. Therefore, in addition to Assumption 2.5, we moreover have the (not restrictive):

Assumption 3.21. Let $\Phi$ be such that $\max_{\alpha, \beta} \|\Phi(t)\| = 1$ and $\forall_{p,q} \max_{\alpha, \beta} \|\phi^p(t)\| = \max_{s} \|\phi^p(s)\|$ (N.B. $\phi^p(t)$ denotes the $p$th column of $\Phi(t)$).

On account of Assumption 3.21 we can estimate $\mathcal{C}_N$ by

$$\mathcal{C}_N \leq \|\Phi(t)\|\|Q^{-1}\|.$$ 

Below, it is shown how sharply $\mathcal{C}_N$ is estimated by $\kappa$.

Property 3.23. Suppose we use the linear norm. Define $R(t)$ by

$$R(t) := \Phi(t) \cdot \text{diag}(\|\phi^1(t)\|^{-1}, \ldots, \|\phi^n(t)\|^{-1}).$$

Then

$$\min_{t} \text{glb} \ (R(t)) \kappa \leq \mathcal{C}_N \leq \kappa.$$ 

Proof. $\max_{t} \|\Phi(t)Q^{-1}\| \leq \max_{t} \{\text{glb} \ (R(t)) \min_{p} \|\phi^p(t)\|\} \kappa \leq \min_{t} \text{glb} \ (R(t)) \kappa. \quad \Box$

Remark 3.25. To appreciate Property 3.23, one should realize that $\text{glb} \ R(t)$ is in fact the condition number of $R(t)$ (cf. Rem. 2.2). In the autonomous case, $R(t)$ may be thought of as the matrix of scaled eigenvectors; except that if they are skew the lower bound in Property 3.23 is substantially smaller than $\kappa$.

We can also show that the scaling assumption 3.21 is optimal regarding the estimate in (3.20), if we use the linear norm, as follows from:

Property 3.26. Let $\mathcal{D}$ denote the class of nonsingular diagonal matrices. For $D \in \mathcal{D}$ let $\kappa(D) = \|\Phi(\alpha)D + M_\beta \Phi(\beta)D^{-1}\|$. Then

$$\min_{D \in \mathcal{D}} \max_{t} \|\Phi(t)\| \kappa(D) = \max_{t} \|\Phi(t)\| \kappa(I) = \kappa.$$ 

Proof. The proof essentially uses the arguments as given in [16, Thm. 23]. \quad \Box

Since the quantity $\kappa$ is easier to handle than $\mathcal{C}_N$ and moreover has turned out to be a fairly sharp estimate of $\mathcal{C}_N$, we will use $\kappa$ rather than $\mathcal{C}_N$ further on.

We finally remark that for separated BC, the quantity $\kappa$ equals

$$\kappa = \left[ \begin{array}{c} \frac{S_\alpha (\Phi(\alpha))}{S_\beta \Phi(\beta)} \end{array} \right].$$
Remark 3.28. For numerical purposes we should in fact consider a somewhat
different condition number. Indeed, in practice we deal rather with a discretized ODE,
where we have (approximate) solutions defined on some grid only. The introduction
of condition numbers of such discrete problems goes similarly to the continuous case,
however. In [12] it is shown that for a sufficiently fine grid one can construct a discrete
fundamental solution that is close in norm to the continuous fundamental solution \( \Phi \);
the conditioning of the discrete and continuous case are almost the same, then. For
stiff problems, however, this is not always true, as is shown in [11].

4. The conditioning of some model problems. As we saw in Example 3.1, the
conditioning may change dramatically if we change the BC. Therefore it is of great
importance to investigate the relation of the conditioning to the BC in general. In
particular we are interested in the question of whether a not-large condition number
might imply that the BC has some special properties. The answer is positive, and as
an interesting application of this we indicate in Example 6.5 why an essential part of
the Godunov–Conte algorithm is stable indeed.

In order to be able to say whether or not a BVP is well conditioned we have to
be more specific about the interpretation of the actual value of \( K \) (which will be used
rather than the number \( \kappa \)). In particular, we want to avoid inexact use of the
terminology “order of.”

One possibility is to consider ODE defined on \([0, \infty)\). By imposing the BC on
the solution \( x \) defined on \([0, T]\) for some \( T > 0 \), we can expect the ill conditioning,
if present, to become more and more pronounced as \( T \to \infty \). Of course, this setting
automatically covers BVP on infinite intervals and in this sense it shows some similarity
with the approach in [6]. Another way to circumvent a sloppy use of the “order-of-
” terminology would be to consider a family of ODE where we can let \( \lambda \) and \(-\mu\) go
to infinity somehow. Such a setting would also cover certain singular perturbation
problems. As the reader may verify quite easily, the latter approach would give similar
results. We do not consider this here, however. We would like to stress that the results
of the theorems below can be used for a qualitative interpretation of finite interval
BVP, i.e., the infinite cases below give a good insight into the nature of the problem
for finite intervals (cf. Example 6.3).

Assumption 4.1. Let the ODE be defined on \([0, \infty)\). For each \( T \) define on \([0, T]\)
a fundamental solution \( \Phi_T \), satisfying the requirements of \( \Phi \) as stated in Assumptions
2.5 and 3.21.

Now consider the family of BC

\[
M_0 x(0) + M_\infty x(T) = b, \quad T > 0 .
\]

Definition 4.3. The family of BVP with BC (4.2) (the BVP, for short) is called
well conditioned if \( \| [M_0 \Phi_T(0) + M_\infty \Phi_T(T)]^{-1} \| = O(1) \), \( T \to \infty \), and ill conditioned if
\( \| [M_0 \Phi_T(0) + M_\infty \Phi_T(T)]^{-1} \|^{-1} = \mathcal{O}(1) \), \( T \to \infty \). The results below will concentrate on how
many independent conditions are needed at both boundary points and also how the
conditioning depends on the relation between the directions of the fundamental
solutions at the boundary points and \( M_0 \) and \( M_\infty \). For the sake of completeness, we
recall that \( k \) indicates the dimension of the dominant subspace, and \( l \), of the dominated
subspace.

Theorem 4.4. The BVP is ill conditioned if \( \text{rank}(M_0) < k \) or \( \text{rank}(M_\infty) < k \).

Proof. Write \( \Phi(t) \) instead of \( \Phi_T(t) \). Suppose, e.g., \( \text{rank}(M_0) = p < l \). Then also
\( \text{rank}(M_0 \Phi(0)) \leq p < l \). Let \( S \) be an \( n \times (n-p) \) matrix of rank \( n-p \) such that \( M_0 \Phi(0)S = 0 \).
Since \( \text{rank}(\Phi^1(T)) = n-l \), there certainly exists an \( r \in \mathbb{R}^{n-p} \), such that
\[ \| S r \| = 1 \text{ and } [\Phi^1(T)|\Phi^2(T)|\varnothing] S r = 0. \] For the vector \( s \in \mathbb{R}^n \), \( s := S r \) we then find
\[ \| [M_0 \Phi(0) + M_\infty \Phi(T)] s \| = \| M_\infty [\Phi^1(T)] s \| \leq \| M_\infty \| \| [\Phi^1(T)] \| = O(e^{-T}). \]
Hence \( \| [M_0 \Phi(0) + M_\infty \Phi(T)]^{-1} \|^{-1} = o(1). \)

**Remark 4.5.** It is not difficult to see that the requirements \( \text{rank } (M_0) \equiv l \) and \( \text{rank } (M_\infty) \equiv k \) (cf. Theorem 4.4) do not imply a well conditioning of the BVP. To see this one should consider a BC where \( M_0 \) and \( M_\infty \) have zeros in the last row.

A more refined result is contained in:

**Theorem 4.6.** The BVP is ill conditioned if
\[ \text{rank } \begin{bmatrix} M_0(0) & \Phi^1(T) \end{bmatrix} < k \quad \text{or} \quad \text{rank } \begin{bmatrix} M_\infty(0) & \Phi^2(T) \end{bmatrix} < k. \]

**Proof.** Like the proof of Theorem 4.4, except for taking \( S \) such that
\[ M_0(0) \Phi^1(T) = 0. \]

**Remark 4.7.** Theorem 4.6 can be interpreted as follows: Suppose precisely \( s \) independent row vectors of \( M_0 \) are orthogonal to \( \phi^{k+1}(0), \ldots, \phi^n(0) \) and precisely \( t \) independent row vectors of \( M_\infty \) are orthogonal to \( \phi^1(T), \ldots, \phi^{n-1}(T) \). Then the problem is ill conditioned if \( \text{rank } (M_0) - s < l \) or \( \text{rank } (M_\infty) - t < k \). In practice, this means that if we have near orthogonality, we can already expect ill conditioning in the finite interval case.

**Remark 4.8.** Note that Theorems 4.4 and 4.6 imply necessary conditions for well conditioning. This will be useful in practical problems which are known to be well conditioned (e.g., on physical grounds).

In order to give a positive result for well conditioning, we should require that a “near orthogonality,” such as was noted in Remark 4.7, does not take place. The next theorem gives a necessary and sufficient condition in case we have a separated BC.

**Theorem 4.9.** Let the BVP have a separated BC, i.e.,
\[ M_0 = \begin{bmatrix} S_0 & \varnothing \\ \varnothing & S_\infty \end{bmatrix} \quad \text{and} \quad M_\infty = \begin{bmatrix} \varnothing \\ S_\infty \end{bmatrix}. \]

Assume either \( \text{rank } (S_0) = l \) or \( \text{rank } (S_\infty) = k \). Then it is well conditioned if and only if
\[ \text{either } \| [S_0(0)|\Phi^1(T)]^{-1} \| = O(1) \quad \text{and} \quad \| [S_\infty(0)|\Phi^1(T)]^{-1} \| = O(1) \]
\[ \text{or } \| [S_0(0)|\Phi^3(T)]^{-1} \| = O(1) \quad \text{and} \quad \| [S_\infty(0)|\Phi^1(T)|\Phi^2(T)]^{-1} \| = O(1). \]

**Proof.** From (3.27) we obtain that the condition number \( \kappa \) equals
\[ \| \begin{bmatrix} S_0(0) & \varnothing \\ \varnothing & \Phi^1(T) \end{bmatrix}^{-1} \|. \]
(a) Let e.g. \( \| [S_0(0)|\Phi^3(T)]^{-1} \| = O(1) \) and \( \| [S_\infty(0)|\Phi^1(T)|\Phi^2(T)]^{-1} \| = O(1) \).

Let \( T \) be sufficiently large, then \( \text{rank } (S_0) = l \). Now write
\[ Q = \begin{bmatrix} S_0(0) \\ \varnothing \end{bmatrix} \Phi^1(T) = \begin{bmatrix} B^1 \\ B^2 \\ \varnothing \end{bmatrix} \left( \begin{array}{c|c|c} \hline & & \hline B^1 & B^2 & C \\ \hline & & \hline D^1 & D^2 & E \end{array} \right) \]
\[ \begin{array}{c|c} \hline & \hline l & n-l \end{array} \]
\[ \begin{array}{c} \hline \hline \hline \hline k \hline \end{array} \]
\[ \begin{array}{c|c} \hline \hline \hline \hline l \hline \end{array} \]
Here $C$ is an $l \times l$-matrix and $[D^1 \mid D^2]$ an $(n-l) \times (n-l)$ matrix, $\|B^1\| = O(e^{-\lambda T})$, $\|E\| = O(e^{\mu T})$. If $B^2 \not= 0$ then there exists some matrix $R$ such that $CR + B^2 = 0$. (N.B. $R = R(T)$!) Now define

$$
\tilde{R} := \begin{bmatrix}
I & \emptyset \\
\emptyset & R & I
\end{bmatrix}
$$

and $\tilde{Q} := Q\tilde{R}$.

Then $\tilde{Q}$ equals $Q$ but for the matrix $B^2$, where $\tilde{Q}$ has zeros, and the matrix $D^2$, where $\tilde{Q}$ has a matrix $\tilde{D}^2 = D^2 + ER$. (N.B. $\|ER\| = O(e^{\rho T})$.) Write with obvious meaning

$$
\tilde{Q} := \begin{bmatrix}
\tilde{B} & \tilde{C} \\
\tilde{D} & \tilde{E}
\end{bmatrix}
$$

Then we have

$$
\tilde{Q}^{-1} = \begin{bmatrix}
o(1) & \tilde{C}^{-1} + o(1) \\
\tilde{D}^{-1} + o(1) & o(1)
\end{bmatrix}
$$

Hence it follows from the assumption that $\|\tilde{Q}^{-1}\| = O(1)$, but since $\|R^{-1}\| = O(1)$ it also follows that $\|Q^{-1}\| = O(1)$.

(b) Suppose the assertion is not true. Let, e.g., $\text{glb}(S_0[\Phi^1(0)]) = o(1)$. Let $y \in \mathbb{R}^l$ be such that $\|S_0[\Phi^1(0)]y\| = \text{glb}(S_0[\Phi^1(0)])\|y\|$. Use the notation $Q := [\begin{bmatrix}\tilde{B} & \tilde{C} \\
\tilde{D} & \tilde{E}\end{bmatrix}$ as above and define $Cy = z^1$; $Ey = z^2$. Then

$$
\text{glb}(Q) = \frac{\|Q[0\ y]\|}{\|y\|} \leq \frac{\|z^1\| + \|z^2\|}{\|y\|} = \text{glb}(C) + O(e^{\rho T}) = o(1).
$$

Hence either $Q^{-1}(T)$ does not exist for almost all $T$ or $\|Q^{-1}\|^{-1} = o(1)$, i.e., the problem is not well conditioned. The other cases go essentially similarly.

**Corollary 4.10.** If the BVP in Theorem 4.9 is well conditioned then $l \leq \text{rank}(S_0) \leq n - k$ and $k \leq \text{rank}(S_\infty) \leq n - l$.

There remains a number of types of ODE that are not represented in the foregoing. We hope that some examples will suffice to show that, e.g., solution spaces with partially increasing and partially decreasing solutions can be dealt with in a similar way.

**Example 4.11.** Consider the ODE $u'' - 4tu' + (4t^2 - 2)u = 0$; this has basis solutions $u(t) = e^{t^2}$ and $u(t) = t e^{t^2}$. Let $\alpha = -T$ and $\beta = T$. After rewriting this ODE in a companion matrix form we obtain a fundamental solution $\Phi$ that, after scaling according to Assumption 3.21, has a $\Phi(0)$ and a $\Phi(T)$ with elements of order unity. As far as the condition number is concerned, we therefore can expect a situation similar to a model problem where $k = l = 0$. 
Example 4.12. Consider the ODE \( u'' + 4tu' + (4t^2 + 2)u = 0 \), where the basis solutions are \( u(t) = e^{-t^2} \) and \( u(t) = t^2 e^{-t^2} \). Again let \( \alpha = -T \) and \( \beta = T \). After rewriting and scaling we now have \( \Phi(0) \) and \( \Phi(T) \) being of the order of \( e^{-T^2} \). This means that we can expect the condition number to be of the order of \( e^{T^2} \), at least.

5. The stability of multiple shooting. A very successful improvement of (single) shooting is multiple shooting (cf. [17, pp. 170 ff.]). For good surveys see, e.g., [2], [8]. Although there seems to be a common opinion about the advantages of this algorithm, the arguments vary quite substantially (cf. [2], [4], [5], [8], [14], [15]). We think that our notion of conditioning can be helpful to clear up some questions about the stability of this algorithm.

We first give a brief description of the method. Let \([\alpha, \beta]\) be divided into subintervals \([t_i, t_{i+1}], i = 0, \cdots, N-1\) (so to \(t_0 = \alpha, t_N = \beta\)). Then for each \(i\) a particular solution \(r_i\) and a fundamental solution \(F_i\) is computed on \([t_i, t_{i+1}]\). Hence for each \(i\) there exists a vector \(v_i\) such that

\[
x(t) = F_i(t)v_i + r_i(t).
\]

For simplicity we assume that \(F_i(t_i) = I\). By matching the relations (5.1) at the points \(t_1, \cdots, t_{N-1}\) we obtain a recurrence relation for the \(v_i\)

\[
A_i v_{i-1} = v_i + g_i, \quad i = 1, \cdots, N-1
\]

where

\[
A_i = F_{i-1}(t_i) \quad \text{and} \quad g_i = r_i(t_i) - r_{i-1}(t_i).
\]

The matrix \(A_i\) in (5.3) is the increment of the fundamental solution going from \(t_{i-1}\) to \(t_i\). In particular we have

\[
F_0(t) = F_i(t)A_i \cdots A_1, \quad t \geq t_i.
\]

If we use (5.1) in the BC (2.7) and add to it the relations (5.2) we obtain the multiple shooting system

\[
\begin{bmatrix}
A_1 & -I & & & \\
& A_2 & -I & & \\
& & & \ddots & \\
& & & & A_{N-1} & -I \\
& & & & & M_{0}A_N
\end{bmatrix}
\begin{bmatrix}
v_0 \\ \vdots \\ v_{N-1}
\end{bmatrix}
= \begin{bmatrix}
g_1 \\ \vdots \\ g_{N-1}
\end{bmatrix} - \begin{bmatrix}
b - M_{\alpha}r_0(t_0) - M_{\beta}r_{N-1}(t_N)
\end{bmatrix}
\]

(N.B. \(A_N := F_{N-1}(t_N)\), cf. (5.3)). The matrix in (5.5) will be denoted by \(P\). Although we actually find incremental matrices \(A_i\) of the discretized problem in practice, we neglect this fact in our discussion. In fact, one may also think of (5.1)–(5.5) as belonging to the discrete problem, cf. Remark 3.28.

In order to study the sensitivity with respect to rounding errors in the final solution, we must specify how the algorithm is carried out. All variants have in common that they use an initial value technique on each subinterval. Therefore, if we denote by \(\xi\) the machine constant, the actually computed \(A_i\) and \(g_i\) may contain errors of the order of

\[
\|A_i\|\xi.
\]

The variant described, e.g., in [3], [17] uses direct forward recursion in order to express \(v_{N-1}\) in \(v_0\), i.e.,
The solution of (5.5) is quite simple now, for we can solve for \( v_0 \). The remaining \( \{v_i\} \) follow from (5.2), and in this way we can determine \( \{x(t_i)\} \). By using (5.7) we may make an error in \( v_{N-1} \) of the order of

\[
(5.8) \quad \max_{t} \left\| \prod_{i=1}^{N-1} A_i \right\| \zeta \leq \max_{0 \leq i \leq N} \left\| F_0(t_i) \right\| \zeta.
\]

In solving for \( v_0 \) we moreover should reckon with an amplification of this error by \( \left\| [M_a + M_0 F_0(\beta)]^{-1} \right\| \). Therefore, roughly speaking, we may expect errors in \( v_0 \) of the order of

\[
(5.9) \quad \max_{t} \left\| F_0(t) \right\| \left\[ [M_a + M_0 F_0(\beta)]^{-1} \right\| \zeta.
\]

The factor in (5.9), in front of \( \zeta \), may be considered as a stability constant ("condition number") for this particular algorithm. It is, however, not the appropriate quantity to indicate the conditioning of the multiple shooting process in general; note moreover that we rejected this factor to estimate \( \zeta_N \) in (3.20). Finally, it is noteworthy that single shooting would give a similar stability constant.

The other variants (cf. [8], [13], [14]) solve the system (5.5) by some stable LU-decomposition of \( P \). For such a method we may therefore expect errors in the \( v_i \) of the order of

\[
(5.10) \quad \left\| P^{-1} \right\| \max_{t} \left\| A_i \right\| \zeta.
\]

Below we now give a realistic estimate for \( \left\| P^{-1} \right\| \). It will turn out that \( \left\| P^{-1} \right\| \) is not large if \( \kappa \) (so \( \zeta \)) is not large. On account of this we thus may conclude that as long as \( \max_{t} \left\| A_i \right\| \zeta \) does not exceed the required tolerance for the global error, the algorithm is numerically stable for a well conditioned problem.

Remark 5.11. In [14] a backward analysis has been given for multiple shooting with separated BC. An essential role in this is played by an estimate for \( \left\| P^{-1} \right\| \); unfortunately, a proof for this estimate is not given in that paper. We would like to remark that our result in Theorem 5.12, below, is different, at least in a qualitative sense. The kind of condition number that appears in [14] seems to be based on the theoretical dependence on the inhomogeneous terms only, whereas ours more explicitly shows the relation to the well conditioning with respect to the BC.

We have:

**Theorem 5.12.** Let \( \| \cdot \| = \| \cdot \|_{\infty} \). For \( i = 0, \ldots, N \) let \( R_i \) be defined by \( \Phi(t_i) = R_i \text{diag}(\|\phi^1(t_i)\|, \ldots, \|\phi^n(t_i)\|) \) (cf. (3.24)). Denote \( \chi = \max_{i,j} \| R_i \| \| R_j^{-1} \| \). Finally, let \( h = \min_i (t_{i+1} - t_i) \). Then

\[
\| P^{-1} \|_{\infty} \leq \chi (\kappa + 1) \left\{ \frac{\gamma_1 e^{-\lambda h}}{1 - e^{-\lambda h}} + \max \left[ \frac{\gamma_2}{1 - e^{-h}}, (N-1) \gamma_3 \right] \right\} + \kappa.
\]

**Proof.** Let \( \Phi \) be a fundamental solution as in Assumptions 2.5 and 3.21. Denote \( \Phi_i = \Phi(t_i) \). Using the \( R_i \) we see that

\[
(a) \quad \begin{bmatrix} B_i & C_i \end{bmatrix} := R_i^{-1} A_i R_{i-1}, \quad B_i \text{ of order } k
\]

is a diagonal matrix. In order to find \( P^{-1} \) we successively solve

\[
(b) \quad PY(q) = E(q), \quad q = 0, \ldots, N-1,
\]

where \( Y(q) \) and \( E(q) \) are block vectors consisting of \( n \) columns. \( E(q) \) contains zeros except for \( E_{q+1}(q) = I \) (by \( E_i(q) \) we denote the \( i \)th block). Now define for such a block vector the notation
\( \hat{Y}_i(q) := R_i^{-1} Y_i(q), \quad i = 0, \ldots, N \) if \( q \leq N - 1 \).

From (b) and (c) we see that \( \{\hat{Y}_i(q)\}_{0}^{N} \) satisfies an inhomogeneous recursion. Therefore we split it into some other suitable particular solution \( \{\bar{U}_i(q)\}_{0}^{N} \), say, and a homogeneous solution \( \{\bar{Z}_i(q)\}_{0}^{N} \), say, in order to find estimates for the \( Y_i(q) \). For simplicity we omit the \( (q) \) for the moment. Let the superscripts 1 and 2 denote a partitioning corresponding to (a), i.e.,

\[
\hat{Y} := \begin{bmatrix} \hat{Y}^1 \\ \hat{Y}^2 \end{bmatrix}.
\]

Then a suitable particular solution \( \{\bar{U}_i\} \) is given by

\[
(d^1) \quad \bar{U}_i^1 = -\left( \prod_{l=1}^{q+1} B_l \right)^{-1} \tilde{E}_q^1, \quad i \equiv q, \quad \bar{U}_i^1 = 0 \quad \text{otherwise};
\]

\[
(d^2) \quad \bar{U}_i^2 = \left( \prod_{l=q+2}^{i} C_l \right) \tilde{E}_q^2, \quad i \equiv q + 1, \quad \bar{U}_i^2 = 0 \quad \text{otherwise}.
\]

The homogeneous solution \( \{\bar{Z}_i\} := \{\hat{Y}_i - \bar{U}_i\} \) then satisfies the BC

\[
(e) \quad M_a R_0 \bar{Z}_0 + M_\beta R_N \bar{Z}_N = S := -M_a R_0 \bar{U}_0 - M_\beta R_N \bar{U}_N.
\]

Using the fact that \( Z_i = \Phi_i \Phi_0^{-1} Z_0 \) we therefore find for \( Y_i \)

\[
(f) \quad Y_i = R_i \bar{U}_i + \Phi_i [M_a \Phi_0 + M_\beta \Phi_N]^{-1} S.
\]

If \( q = N - 1 \), then \( \{\hat{Y}_i(q)\} \) just satisfies a homogeneous recursion and we immediately obtain

\[
(g) \quad Y_i = \Phi_i [M_a \Phi_0 + M_\beta \Phi_N]^{-1}.
\]

The required estimate then follows from (f) and (g). To see this, denote for short:

\[
\rho_i = \gamma_i \exp[-\lambda hi], \quad \nu_i = \max(\gamma_2 \exp[\mu hi], \gamma_3), \quad \eta_i = \rho_i + \nu_{N-i-1}.
\]

Then for \( q < N - 1 \):

\[
\begin{bmatrix} \|Y_0(q)\|_{\infty} \\ \vdots \\ \|Y_{N-1}(q)\|_{\infty} \end{bmatrix} \leq \chi \begin{bmatrix} p_{q+1} \\ \vdots \\ \nu_0 \\ \vdots \\ \nu_{N-q-2} \end{bmatrix} + \chi \kappa \begin{bmatrix} \eta_{N-q-1} \\ \vdots \\ \eta_{N-q-1} \end{bmatrix}
\]

and for \( q = N - 1 \):

\[
\|Y_0(N-1)\|_{\infty}, \ldots, \|Y_0(N-1)\|_{\infty}^T \leq \kappa.
\]

An estimate for \( \|P^{-1}\|_{\infty} \) will be given by considering the maximal row sum in the matrix with elements \( (\|Y_i(q)\|_{\infty}) \). Since

\[
\sum_{j=1}^{N} \eta_j \equiv \sum_{j=0}^{N-2} \nu_j + \sum_{j=1}^{N-1} \rho_j,
\]

\[
\sum_{j=0}^{N-2} \nu_j \equiv \max \left[ \gamma_2 \frac{e^{-\mu h}}{1-e^{-\mu h}}, (N-1)\gamma_3 \right], \quad \sum_{j=1}^{N-1} \rho_j \equiv \gamma_1 \frac{e^{-\lambda h}}{1-e^{-\lambda h}}
\]

the estimate immediately follows. \( \square \)

**Corollary 5.13.** If \( (k + l) < n \) (so there are solutions that do not increase nor decrease) then \( \|P^{-1}\|_{\infty} \leq N \kappa \chi \).
COROLLARY 5.14. If \((k + l) = n\) and \(\lambda h, -\mu h\) are not small, then \(\|P^{-1}\|_{\infty} \leq \kappa\chi\). If, however, \(\lambda h\) or \(-\mu h\) is small, then \(\|P^{-1}\|_{\infty} \approx N\kappa\chi\) again.

Remark 5.15. The result in Corollary 5.13 regarding the factor \(N\) is rather sharp, as can be seen from a \(P\) where the matrices \(A_i\) and \(M_\beta\) are equal to \(I\) and where \(M_\alpha = 0\). Here we have \(\|P^{-1}\|_{\infty} = N\).

Remark 5.16. Since the norms of the incremental matrices \(\|A_i\|\) are usually significantly larger than 1, we see that \(\|P\| \approx \max \|A_i\|\). Therefore we obtain from (5.10) that \(\|P^{-1}\|\|P\|\) gives a good idea of the absolute rounding error amplification.

6. Examples. In this section we give a number of examples to illustrate the results of \(\S\) 4 and 5. In order to eliminate discretization errors, we constructed the multiple shooting matrices from analytically known fundamental solutions. The norm is assumed to be the supremum norm \(\|\cdot\|_{\infty}\).

Example 6.1. Consider the ODE

\[
\frac{dx}{dt} = \begin{bmatrix}
1 - 2 \cos 2t & 0 & 1 + 2 \sin 2t \\
0 & 2 & 0 \\
-1 + 2 \sin 2t & 0 & 1 + 2 \cos 2t
\end{bmatrix} x
\]

on \([0, \pi]\). A fundamental solution \(F_0\) (with \(F_0(0) = I\)) is given by

\[
F_0(t) = \begin{bmatrix}
\sin t & 0 & -\cos t \\
0 & 1 & 0 \\
\cos t & 0 & \sin t
\end{bmatrix} \cdot \text{diag}(e^{3t}, e^{2t}, e^{-t}).
\]

Assume \(M_0 = M_\pi = I\). For the scaled fundamental solution \(\Phi\) we obtain

\[
\Phi(0) = \begin{bmatrix}
0 & 0 & 0 \\
0 & e^{-2\pi} & 0 \\
e^{-3\pi} & 0 & 0
\end{bmatrix}, \quad \Phi(\pi) = \begin{bmatrix}
0 & 0 & -e^{-\pi} \\
0 & 1 & 0 \\
-1 & 0 & 0
\end{bmatrix}.
\]

Hence

\[
M_0\Phi(0) + M_\pi\Phi(\pi) = \begin{bmatrix}
0 & 0 & -1 + e^{-\pi} \\
0 & 1 + e^{-2\pi} & 0 \\
-1 + e^{-3\pi} & 0 & 0
\end{bmatrix}.
\]

We thus find \(\kappa = [-1 + e^{-\pi}]^{-1} \approx 1.05\). It can also be checked that \(\mathcal{C}N \approx 1.05\). This sharpness of the estimate \(\kappa\) is confirmed by 3.23, from which we deduce that \(\kappa/\sqrt{2} \leq \mathcal{C}N \leq \kappa\). Therefore, we should call this BVP well conditioned. Using the notation for the dimension of the dominant solution space, \(k\), and the subdominant solution space, \(l\), as in Assumption 2.5, we apparently have \(k = 2\) and \(l = 1\). From the well conditioning, we may conclude that rank \((M_0) \geq 1\) and rank \((M_\alpha) \geq 2\), as follows from Theorem 4.4. This is trivially true, of course. We also computed the multiple shooting matrix. For this we used \(N + 1\) equally spaced points \(t_0, \cdots, t_N\). For checking the estimate in Theorem 5.12 it is important to note that \(\chi, \gamma_1\) and \(\gamma_2\) are about \(\sqrt{2}\). Moreover, for \(N\) not too small, we see that max \((e^{-\lambda\pi/N}/(1 - e^{-\lambda\pi/N}), 1 - e^{-\mu\pi/N}) \approx N/\pi\) (N.B. \(\lambda = 2, \mu = -1\)). Hence as an estimate for \(\|P^{-1}\|\) we would find \(\|P^{-1}\| \approx 4N/\pi\). In Table 6.1 we have given the actually computed values; they seem to agree with this estimate apart from a factor \(\approx 4\).
As we said in Remark 5.16, $\|P\|$ may be considered to be indicative of the maximal incremental growth. Therefore we have also given the condition number of $P$ in Table 6.1, which can be used to estimate the error in (5.10). Note that the values for $N = 1$ correspond to single shooting.

**Example 6.2.** Suppose we have an ODE like (6.1a) but now with

$$F_0(t) = \begin{bmatrix} \sin t & 1 & -\cos t \\ 0 & 1 & 0 \\ \cos t & 0 & \sin t \end{bmatrix} \cdot \text{diag} \left( e^{20t}, e^{19t}, e^{-18t} \right).$$

We now have $\kappa = 1.0$ (up to 26 decimals). In Table 6.2 we have given values for the corresponding multiple shooting matrix. Since we have a strong dichotomy now, we almost do not see a dependence on $N$ (cf. Cor. 5.14). For smaller $N$ the incremental values are very large now. If we have a 16-digit mantissa, we cannot hope to obtain more than 7 accurate decimals for $N = 3$, for example; note that single shooting will then completely fail.

**Example 6.3.** Again consider a similar ODE, now with a fundamental solution $F_0$ given by

(a) $$F_0(t) = \begin{bmatrix} \sin t & 0 & -\cos t \\ 0 & 1 & 0 \\ \cos t & 0 & \sin t \end{bmatrix} \cdot \text{diag} \left( e^{3t}, 1, e^{-t} \right).$$

and a separated BC, where

$$S_0 = (1 \ 0 \ 0), \quad S_\pi = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$  

In order to investigate the conditioning we consider

(b) $$Q = \begin{bmatrix} S_0 & \Phi(0) \\ S_\pi & \Phi(\pi) \end{bmatrix} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix}. $$

Hence $\kappa = 1$. We also find

(c) $$\|\{S_0[\Phi^3(0)]\}^{-1}\| = 1 \quad \text{and} \quad \|\{S_\pi[\Phi(\pi)\Phi^2(\pi)]\}^{-1}\| = 1. $$

The results in (b) and (c) are in agreement with a finite interpretation of Remark 4.8. For this, one should replace a statement like “$\|a(T)\| = O(1), T \to \infty$” by “$a(T)$ not large.” Since we see from (a) that $k + l < n$, we can expect an $O(N)$ behavior of
This is confirmed by Table 6.3.

\[ \|P^{-1}\| \]

<table>
<thead>
<tr>
<th>( N )</th>
<th>( |P^{-1}| )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>24</td>
<td>24</td>
</tr>
</tbody>
</table>

**Example 6.4.** Let the ODE be as in (6.1a), but now with BC

\[ M_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad M_{\pi} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}. \]

Since \( \text{rank}(M_0) = 2 \) (\( k = l \)) and \( \text{rank} M_{\pi} = 2 \) (\( \geq k \)), we cannot conclude ill conditioning on account of Theorem 4.4. Similarly, since \( \text{rank}(M_0[\Phi^3(0)]) = 1 \) and \( \text{rank}(M_{\pi}[\Phi^3(\pi)]) = 2 \), we cannot conclude ill conditioning on account of Theorem 4.6. Finally, since moreover \( \text{rank}(M_0) + \text{rank}(M_{\pi}) > 3 \) and even \( \text{rank}(M_0 + M_{\pi}) = 3 \) we might hope to have a well conditioned problem. Nevertheless we obtain

\[ (a) \quad Q = \begin{bmatrix} -1 & 0 & -1 \\ 0 & 1 & 0 \\ e^{-3\pi} & 0 & 0 \end{bmatrix}, \quad \text{so } \kappa = 1 + e^{3\pi} \approx 1.2 \cdot 10^5. \]

The values for \( \|P^{-1}\| \) in Table 6.4 confirm this. The result in (a) also shows that an approach such as is suggested in [5, e.g., Thm. 3.3] is not very relevant as far as the estimates for the “condition number” \( \|M_0 + M_{\pi}\| \|M_0 + M_{\pi}\|^{-1} \) are concerned. Also the other quantity which appears there, in our notation \( \max_{t \in [0, \pi]} \|F_0(t)\| \|F_0(t)\|^{-1} \), does not really matter in a realistic bound for error amplification, as follows from our discussion in § 5. Note that the latter quantity \( =e^{4\pi} \) in this example; if we would have had a decreasing solution behaving like \( e^{\mu t}, \mu < 0 \), then this quantity would have been \( =e^{3\pi - \mu \pi} \). Nevertheless \( \kappa \) would remain the same (and therefore \( \|P^{-1}\| \) would not change drastically). It goes without saying that \( \|P\| \) would be almost independent of the magnitude of \( \mu \) (cf. Remark 5.16).

**Example 6.5.** Theorems 4.4, 4.6 and 4.9 also provide useful arguments to explain the stability of certain algorithms. As an example, we consider the Godunov–Conte algorithm [1], [14], [15]: In order to have stability, we e.g., have to prove that an essential part of the method, viz. the backward recursion, cf. [15, Eq. (11)], is stable. For this, assume that \( l + k = n \) and that the BVP (with separated BC) is well conditioned. Then it follows from Theorem 4.4 that \( \text{rank}(S_0) = l \). This means that the orthogonal complement of the space spanned by the rows of \( S_0 \) must be \( k \)-dimensional. The idea of the algorithm now is to compute \( k \) independent basis solutions, in matrix notation \( F_0^1 \), say, such that \( S_0 F_0^1(\alpha) = 0 \), i.e., \( \text{span}(F_0^1(\alpha)) \subset \text{span}(S_0^T) \). On the other hand, we derive from Theorem 4.9 that no column of \( \Phi_0^3(\alpha) \) (i.e., an initial value of
a dominated solution) can be (almost) orthogonal to the space spanned by the rows of $S_0$. Therefore we conclude that $F_0^2(\alpha)$ must consist of dominant solutions. At the shooting points we only reorthogonalize the columns of these solution vectors, whence we see that $\text{span}(F_0^1(t)) = \text{span}(F_0^2(t))$. On account of the analysis as given in [9], [10], it can then be deduced that the corresponding incremental matrices define a recursion that is indeed stable in the backward direction. For a different kind of analysis, see [14].

REFERENCES